AMENDMENTS THE CLAIMS

The Assignee submits below a complete listing of the current claims, including marked-up claims with insertions indicated by underlining and deletions indicated by strikeouts and/or double bracketing. This listing of claims replaces all prior versions and listings of claims in the application:

1. (Currently amended) A program storage device readable by a machine, tangibly embodying a program of instructions executable by the machine to perform a method for speech synthesis that allows user specified pronunciations, the method comprising:

~~aligning a text string comprising a plurality of words and phenomes and a user specified spoken audio signal corresponding to a desired pronunciation of the text string;~~

extracting prosodic parameters from ~~said spoken~~ an audio signal corresponding to a pronunciation of a text string by a user;

extracting duration parameters by aligning the audio signal with the text string;

automatically generating ~~a marked-up text corresponding to the spoken audio signal~~ at least one text-to-speech (TTS) input using the prosodic parameters and the duration parameters, wherein the at least one TTS input is formatted for use in synthesizing speech from the text string; and

generating a synthetic speech waveform using the ~~marked-up text~~ at least one TTS input.

2-3. (Cancelled)

4. (Currently amended) The program storage device of claim 1, wherein the instructions for aligning comprise instructions for segmenting ~~said spoken~~ the audio signal into time-segmented regions, wherein each time-segmented region is mapped to a corresponding phoneme.

5. (Previously Presented) The program storage device of claim 1, wherein the alignment is performed using a Viterbi alignment process.

6. (Canceled)

7. (Currently amended) The program storage device of claim 1, wherein the instructions for automatically generating ~~a marked-up text~~ at least one TTS input comprise instructions for directly specifying at least one portion of the duration ~~contours~~ parameters and/or the prosodic parameters as attribute values for mark-up elements.

8. (Currently amended) The program storage device of claim 1, wherein the instructions for automatically generating ~~a marked-up text~~ at least one TTS input comprise instructions for assigning abstract labels to at least one portion of the duration ~~contours~~ parameters and/or the prosodic parameters to generate a high-level markup.

9. (Currently amended) The program storage device of claim 1, wherein the ~~marked-up text~~ at least one TTS input is generated using SSML (speech synthesis markup language).

10. (Currently amended) The program storage device of claim 1, further comprising instructions for processing phonetic content of the ~~spoken~~ audio signal to generate the synthetic speech waveform having a desired pronunciation.

11. (Currently amended) A method for speech synthesis that allows user specified pronunciations, the method comprising:

~~aligning a text string comprising a plurality of words and phenomes and a user specified spoken audio signal corresponding to a desired pronunciation of the text string;~~

extracting prosodic parameters from ~~said spoken~~ an audio signal corresponding to a pronunciation of a text string by a user;

extracting duration parameters by aligning the audio signal with the text string;

automatically generating ~~a marked-up text corresponding to the spoken audio signal~~ at least one text-to-speech (TTS) input using the prosodic parameters and the duration parameters, wherein the at least one TTS input is formatted for use in synthesizing speech from the text string; and

generating a synthetic speech waveform using the ~~marked-up text~~ at least one TTS input.

12-13. (Canceled)

14. (Currently amended) The method of claim 11, wherein aligning comprises extracting acoustic feature data from the ~~spoken~~ audio signal and time-aligning the ~~spoken~~ audio signal to the text string using the acoustic feature data.

15. (Previously Presented) The method of claim 11, wherein aligning is performed using a Viterbi alignment process.

16. (Canceled)

17. (Currently amended) The method of claim 11, wherein automatically generating ~~a marked-up text~~ at least one TTS input comprises directly specifying at least one portion of the duration ~~contours~~ parameters and/or the prosodic parameters as attribute values for mark-up elements.

18. (Currently amended) The method of claim 11, wherein automatically generating ~~a marked-up text~~ at least one TTS input comprises assigning abstract labels to at least one portion of the duration ~~contours~~ parameters and/or the prosodic parameters to generate a high-level markup.

19. (Currently amended) The method of claim 11, wherein the ~~marked-up text~~ at least one TTS input is generated using SSML (speech synthesis markup language).

20. (Currently amended) The method of claim 11, further comprising processing phonetic content of the ~~spoken~~ audio signal to generate the synthetic speech waveform having a desired pronunciation.

21. (Currently amended) A text-to-speech (TTS) system that allows user specified pronunciations, comprising:

a prosody analyzer for determining prosodic parameters of ~~a spoken~~ an audio signal corresponding to a ~~desired~~ pronunciation by a user of an input text string and automatically generating ~~a marked-up text~~ at least one TTS system input corresponding to the ~~spoken~~ audio signal using the prosodic parameters, wherein the prosody analyzer comprises:

a prosodic parameter extraction module for extracting the prosodic parameters from the audio signal,

an alignment module for extracting duration parameters by aligning the input text string with the ~~spoken~~ audio signal,

~~a prosodic parameter extraction module for determining prosodic parameter information for the spoken audio signal;~~ and

a conversion module for ~~including markup in the input text string~~ generating the at least one TTS system input using the prosodic parameters ~~information to generate marked-up text~~ and the duration parameters; and

a TTS system for generating a synthetic waveform using the ~~marked-up text~~ at least one TTS system input.

22. (Currently amended) The system of claim 21, further comprising a user interface that enables a user to input the ~~spoken~~ audio signal and the input [[a]] text string corresponding to the ~~spoken~~ audio signal.

23. (Currently amended) The system of claim 21, wherein the prosody analyzer processes phonetic content of the ~~spoken~~ audio signal to generate the synthetic waveform having a desired pronunciation.

24-28. (Canceled)

29. (Currently amended) The program storage device of claim 33, wherein extracting acoustic feature data from ~~said spoken~~ the audio signal comprises digitizing the ~~spoken~~ audio signal into a

set of frames and transforming the digitized ~~input waveforms~~ audio signal into a set of feature vectors on a frame-by-frame basis.

30. (Currently amended) The program storage device of claim 29, wherein transforming the digitized ~~input includes~~ audio signal comprises producing a 24-dimensional cepstra feature vector for every 10ms of the ~~spoken~~ audio signal, concatenating frames to the left and to the right of a current frame to augment a current cepstral vector, and reducing each augmented cepstral vector to a 60-dimensional feature vector using linear discriminant analysis.

31. (Currently amended) The method of claim 34, wherein extracting acoustic feature data from ~~said spoken~~ the audio signal comprises digitizing the ~~spoken~~ audio signal into a set of frames and transforming the digitized ~~input waveforms~~ audio signal into a set of feature vectors on a frame-by-frame basis.

32. (Currently amended) The method of claim 31, wherein transforming the digitized ~~input includes~~ audio signal comprises producing a 24-dimensional cepstra feature vector for every 10ms of the ~~spoken~~ audio signal, concatenating frames to the left and to the right of a current frame to augment a current cepstral vector, and reducing each augmented cepstral vector to a 60-dimensional feature vector using linear discriminant analysis.

33. (Currently amended) The program storage device of claim 1 wherein the method further comprises extracting acoustic feature data from ~~said spoken~~ the audio signal and wherein the aligning further comprises outputting a set of duration contours.

34. (Currently amended) The method of claim 11 further comprising extracting acoustic feature data from ~~said spoken~~ the audio signal and wherein the aligning further comprises outputting a set of duration contours.

35. (Currently amended) The system of claim 21 wherein the prosody analyzer further comprises an acoustic feature extraction module that extracts acoustic feature data from ~~said spoken~~ the audio signal and wherein the alignment module uses said acoustic feature data to perform the aligning.